Summary of the Results of Phase I ELFT Testing

24 September 2007

Evaluation of Latent Fingerprint Technology (ELFT) is a NIST project for evaluating automated one-to-many latent fingerprint search technology. ELFT was planned as a multi-phased project. Phase I was designed as a "proof of concept" test for one-to-many automated latent searches. Software was submitted to NIST for testing in the form of *Software Development Kits*, or SDKs. These were installed and run on NIST computers, using NIST data. Additional information on ELFT and Phase I may be found on the ELFT website, particularly under the CONOPS and the API links
↳ http://fingerprint.nist.gov/latent/elft07/

Phase I testing was conducted at NIST during June and July of 2007.
A brief summary of the ELFT Phase I Testing results is presented below.

There were ten participants in Phase I. Each participant was allowed to submit two SDKs, a *primary* and a *secondary*. The primary was intended to be the "best effort" -- optimized for accuracy -- while the secondary could be an alternate implementation, possibly faster but less accurate. A total of sixteen SDKs were received by NIST, ten *primary* and six *secondary*. Prior to being run on the Phase I dataset each SDK had to pass Validation Testing. The purpose of Validation Testing was to ensure that each SDK, when installed at NIST, could reproduce the participant's submitted candidate list. Although some problems were encountered, all sixteen eventually passed Validation Testing. See ↳
http://fingerprint.nist.gov/latent/elft07/validation_testing.pdf

The Phase I dataset consisted of 100 search images, drawn from a number of sources. The background consisted of 1000 (rolled-impression) ten-prints, for a total of 10,000 fingers. The selected latents were of graduated difficulty, though none were intentionally drawn from the lowest quality grades. The latent search images were predominantly scanned at 500 ppi resolution, though about twenty latents scanned at 1000 ppi were included. The size and aspect ratio of images was intentionally varied, though restricted to a maximum size of 1000 x 1000 pixels. Search images varied from "very clean," to a moderate amount of background clutter, including: writing, smudges, and ancillary fingerprints within the image space. In summary, the Phase I dataset was much more varied and difficult than the validation dataset.

The published ground rules for Phase I stipulated that only the number, and not the names, of participants were to be published. Also, only aggregate (average) performance results were to be published.

Of the 16 SDKs completing Validation Testing, one experienced serious difficulties and was unable to complete Phase I. Figure 1 shows the aggregate search accuracy from the 15 SDKs which completed Phase I (including both primary and secondary SDKs) in the form of "box and whiskers plots." The M_1 metric measures the fraction in top position (only), while the M_2 metric gives partial credit for mates appearing in lower positions (1/2 point for second place, 1/3 for third place, etc.). For both M_1 and M_2 the range of possible values is 0 to 1.0. Seven search images had inadequate overlap with their corresponding rolled-impression, and were treated as "no mate" cases. They were excluded when computing the aggregate performance score.

As will be seen from Figure 1, the median score for M_1 was 0.64 (mean = 0.59), while for M_2 the median was 0.66 (mean = 0.61). Note that the difference between the two metrics, M_1 and M_2, is generally small, about 2% in this case, confirming that most hits are in first place, or else do not appear on the candidate list.

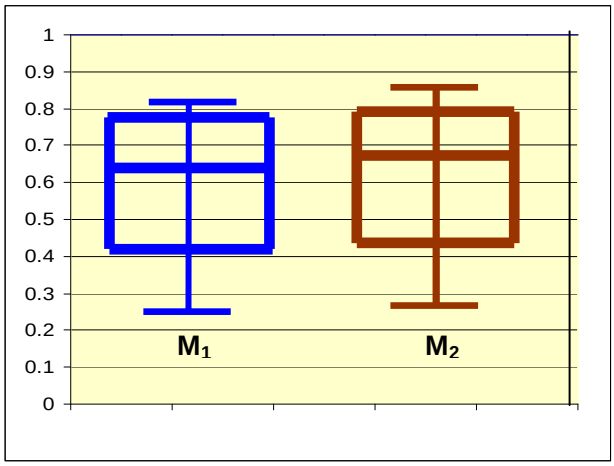


Figure 1 --Box-and-Whiskers Plot of Accuracy

Figure 2 compares the Phase I performance with that obtained during validation testing. Data presented is the average over all 15 SDKs. It will be seen that the average performance for Phase I is approximately 0.6; whereas the average for the Validation Dataset Search Prints is approximately 0.67 when searching against the Phase I background of 1000, and 0.8 when searching against the Validation Dataset Background of 100.

Fig. 2 -- Aggregate (Average) Performance Compared with Validation Set

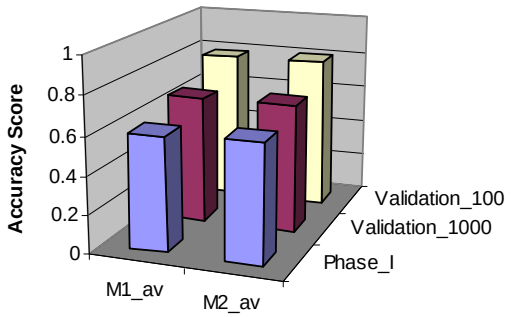


Figure 3 provides a breakdown of performance by primary and secondary SDK. It was intended that the primary SDK be the dominant one, taking more time for execution but producing better results. The secondary SDK was intended to trade performance for speed of execution. In general it was found that the primary SDK did outperform the secondary ones. But because not all participants presented secondary ones, and those that did so tended to have higher-performing SDKs, the average of the secondary SDKs is actually higher than that of the primary, though the difference is not large. If we restrict the results to *participants that submitted both* (a primary and secondary), the average difference is 0.05 in favor of the primary.

Fig. 3 -- Phase_I Aggregate (Average) Performance

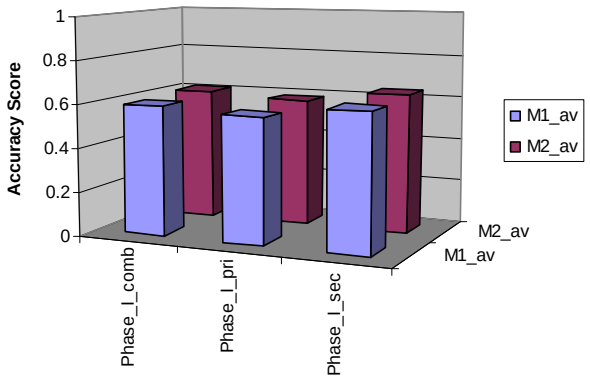


Figure 4 compares the average performance with that of the single best SDK. Also shown are, (1) the percentage of searches identified ("hit") by all participants, about 11%, as well as (2) the percentage hit by at least one SDK. The last category does not require the hit be in first place – only on the candidate list.

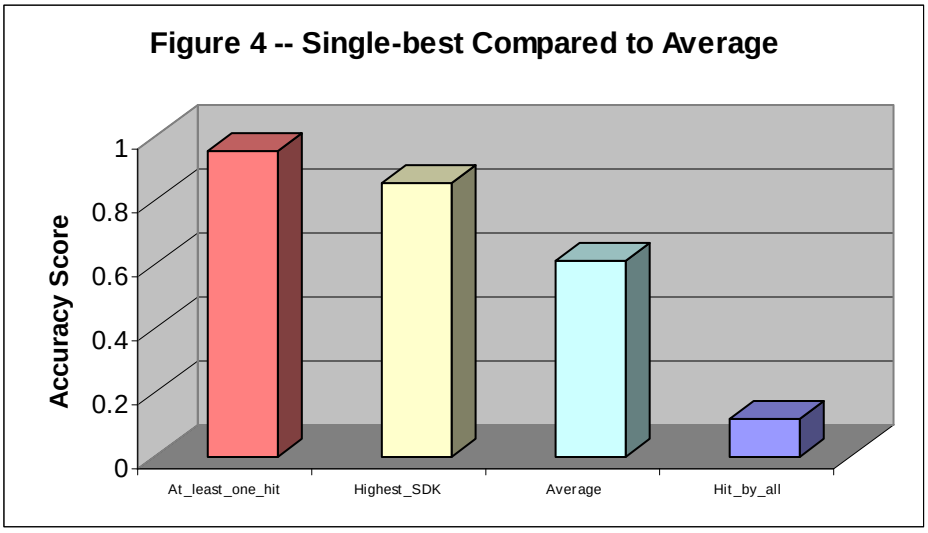


Figure 5 shows an example of a latent search which was "hit" by all SDKs. This search is considered moderately difficult, as it contains a high clutter level, including secondary fingerprints. Despite these challenges all SDKs successfully identified the mate.

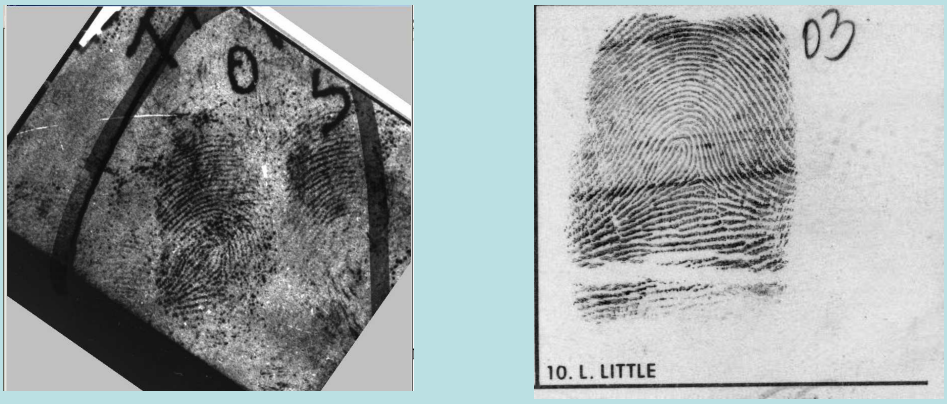


Figure 5 -- Example of a moderately difficult search (left) successfully "hit" (identified) by all SDKs. Mate is at right.

Figure 6 presents a more challenging example. Only about a quarter were successful with the search scanned at 500 ppi, but over half were successful at 1000 ppi.

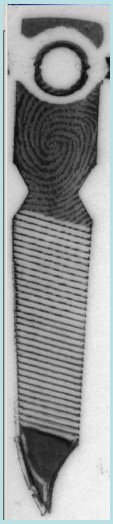


Figure 6 -- Example of a more difficult search (left). Most SDKs missed this at 500 ppi; but a majority were successful at 1000 ppi. Mate is at right.

Figure 7 shows one of the more difficult searches, and was missed by all SDKs.

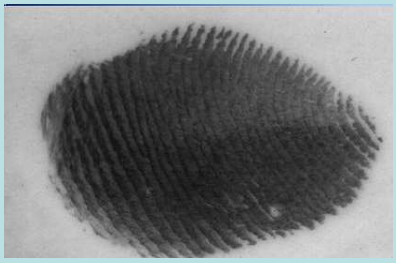


Figure 7 -- Example of a difficult search (left) missed by every SDK. Mate is at right.

Figure 8 presents aggregate processing (execution) times. These times include both enrollment and search. It will be seen that execution times for the primary SDKs is generally larger than for the secondary SDKs, as was expected. However, minimum primary execution time was achieved by a participant who did not submit a secondary.

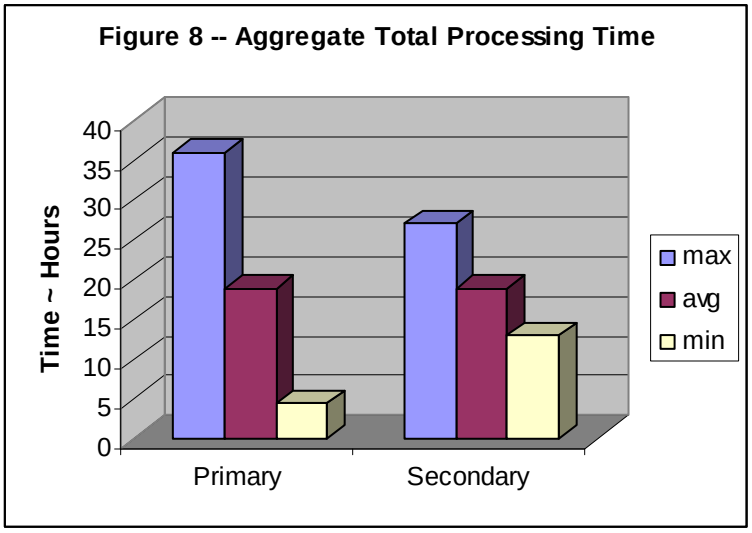


Figure 9 shows data similar to Figure 8, but this time the background enrollment time is excluded. The times shown are therefore the sum of the enroll latent time and the search background time.

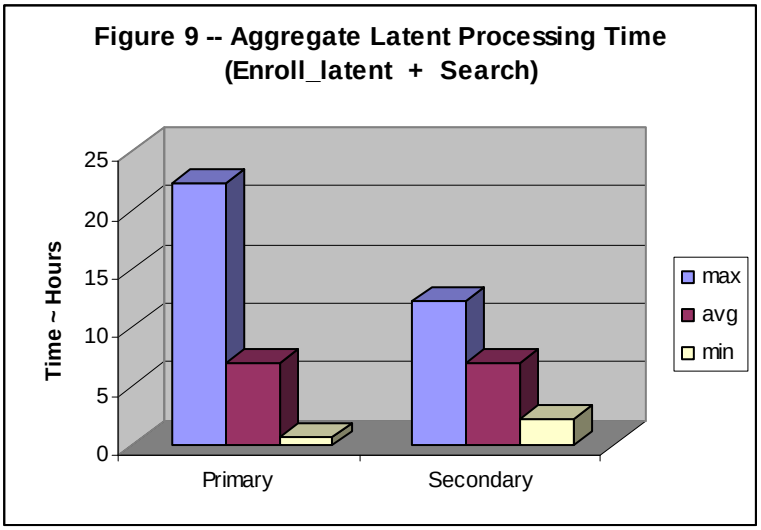
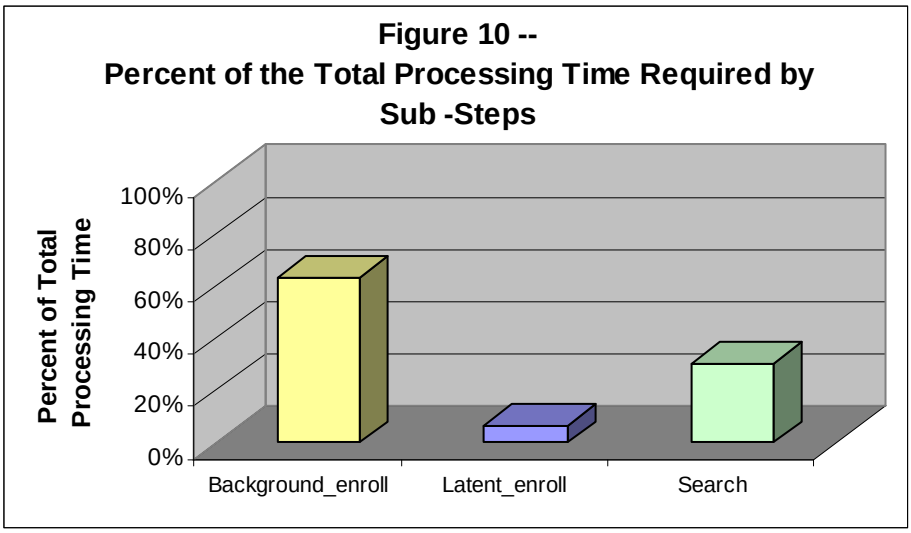


Figure 10 breaks the total execution time down by the major processing steps: enroll background; enroll latent; and search.



It is planned to augment this report in the near future by including an analysis of performance by image-type/image-quality. Also planned is an analysis of the efficacy of the normalized score (probability of true match). The latent and standards communities have expressed significant interest in the development of this type of measure.